Machine Learning Research
2026; 11(1): 8-21
http://www.sciencepublishinggroup.com/j/mlr
doi: 10.11648/j.mlr.20261101.12

Data Science and Machine Learning for Cyber Intrusion Detection: A Systematic Review

Yali Ren*

School of Computer Science, Georgia Institute of Technology, Atlanta, the United States

Email address:
yren78@gatech.edu (Yali Ren)
*Corresponding author

To cite this article:
Yali Ren. (2026). Data Science and Machine Learning for Cyber Intrusion Detection: A Systematic Review. *Machine Learning Research, 11*(1), 8-21. https://doi.org/10.11648/j.mlr.20261101.12

Received: 28 February 2026 ; **Accepted:** 11 March 2026 ; **Published:** 18 March 2026

Abstract: The escalating sophistication and volume of cyberattacks have driven an urgent demand for intelligent Intrusion Detection Systems (IDS) that leverage Data Science (DS) and Machine Learning (ML). Despite rapid advances, existing reviews often focus narrowly on specific aspects without integrating the full data science and machine learning lifecycle. This paper presents a systematic review of DS and ML applications in cyber intrusion detection, covering 153 studies published from 2009 to 2025. The review systematically surveys benchmark datasets, data preprocessing and feature engineering techniques, classical ML and Deep Learning (DL) models, ensemble and hybrid strategies, class imbalance handling, and evaluation methodologies. A unified four-axis taxonomy is proposed to classify the literature, including learning strategy, imbalance handling, explainability level, and deployment context. A quantitative meta-analysis reveals that UNSW-NB15 and CIC-IDS2017 dominate at 71% combined dataset usage, deep learning represents 40% of algorithmic approaches, and only 34% of studies report per-class recall for minority attack types. Nine technically grounded research gaps are identified, spanning preprocessing standardization, cross-dataset evaluation, minority-class recall optimization, adversarial robustness, online and edge deployment, explainability for Security Operations Center (SOC) operations, federated learning, transformer and Large Language Model (LLMs) application, and zero-shot adaptation. The review further identifies eight emerging trends including attention-based and transformer architectures, LLMs, Graph Neural Networks (GNNs), federated and privacy-preserving learning, adversarial robustness, Explainable AI (XAI), zero-shot and few-shot detection, and Internet of Things (IoT) edge-based IDS. A seven-stage actionable architecture is proposed that integrates adaptive preprocessing, contrastive feature learning, recall-aware ensemble detection, XAI decision support, continual learning, and federated aggregation. This review provides researchers and practitioners with a structured roadmap for advancing the next generation of intelligent cyber intrusion detection systems.

Keywords: Intrusion Detection, Machine Learning, Deep Learning, Data Science, Explainable AI

1. Introduction

Cybercrime has become one of the biggest challenges that impair global business success [1]. The proliferation of Internet of Things (IoT) devices and migration to cloud and edge computing have dramatically expanded the attack surface available to threat actors [2, 3]. Cyberattacks have evolved from simple denial-of-service floods to sophisticated intrusions involving ransomware, zero-day exploits, and advanced persistent threats (APTs) [4], which enables traditional cyber intrusion defenses to be increasingly ineffective [5].

Intrusion Detection Systems (IDS) serve as a critical line of defense by monitoring network traffic for signs of malicious behavior. The past decade has seen a transformative shift toward data-driven approaches leveraging data science (DS) and machine learning (ML) to automatically learn discriminative patterns from network traffic [6, 7]. Data Science and Machine Learning based IDS manifests potential for detecting fast mutating and polymorphic threats that evade conventional cyber threat methods [8]. The last ten years have

witnessed rapid advances across a wide range of dimensions, which include but are not limited to data preprocessing pipelines, diverse ML and deep learning (DL) architectures, ensemble and hybrid strategies, novel approaches to class imbalance, and the emergence of transformer-based models, graph neural networks (GNNs), federated learning, and explainable Artificial Intelligence (XAI) [2, 3]. Concurrently, improved benchmark datasets have opened the possibility for more rigorous assessments of IDS performance [9, 10].

Despite these advances, existing reviews often focus narrowly on specific aspects, such as deep learning architectures, feature selection methods, or individual datasets, with less emphasis on integrating the full data science lifecycle [11–13]. There is a need for a comprehensive review that integrates the full data science lifecycle that is from raw data preprocessing through model deployment while identifying the concrete and actionable research gaps.

This paper addresses that gap with a systematic review of data science and machine learning applications in IDS in the last ten years. Figure 1 presents a comprehensive concept map of seven major Data Science and Machine Learning areas in cyber intrusion detection. The specific contributions of this paper include the following aspects: 1). Proposed a unified four-axis taxonomy and made a quantitative meta-analysis of IDS; 2). Presented surveys of preprocessing, feature engineering, classical ML, deep learning, and ensemble strategies from the data science lifecycle perspective; 3). Conducted a dedicated analysis of imbalanced learning and the minority attack collapse phenomenon; and 4). Identified nine research gaps and emerging trends with an actionable next-generation IDS architecture.

The remainder of this paper is organized as follows. Section 2 presents intrusion detection paradigms, benchmark datasets, and a unified taxonomy. Section 3 reviews data science and feature engineering techniques in IDS. Section 4 surveys classical machine learning methods in IDS. Section 5 examines deep learning methods in IDS. Section 6 investigates the ensemble and hybrid strategies in IDS. Section 7 provides a quantitative meta-analysis of the IDS literature and analyzes the strategies and methods for handling class imbalance. Section 8 discusses the challenges and research gaps. Section 9 explores the emerging trends and new paradigms of data science and machine learning in IDS, and proposes a cutting-edge actionable architecture for IDS. Section 10 concludes the paper and outlines recommendations for future research.

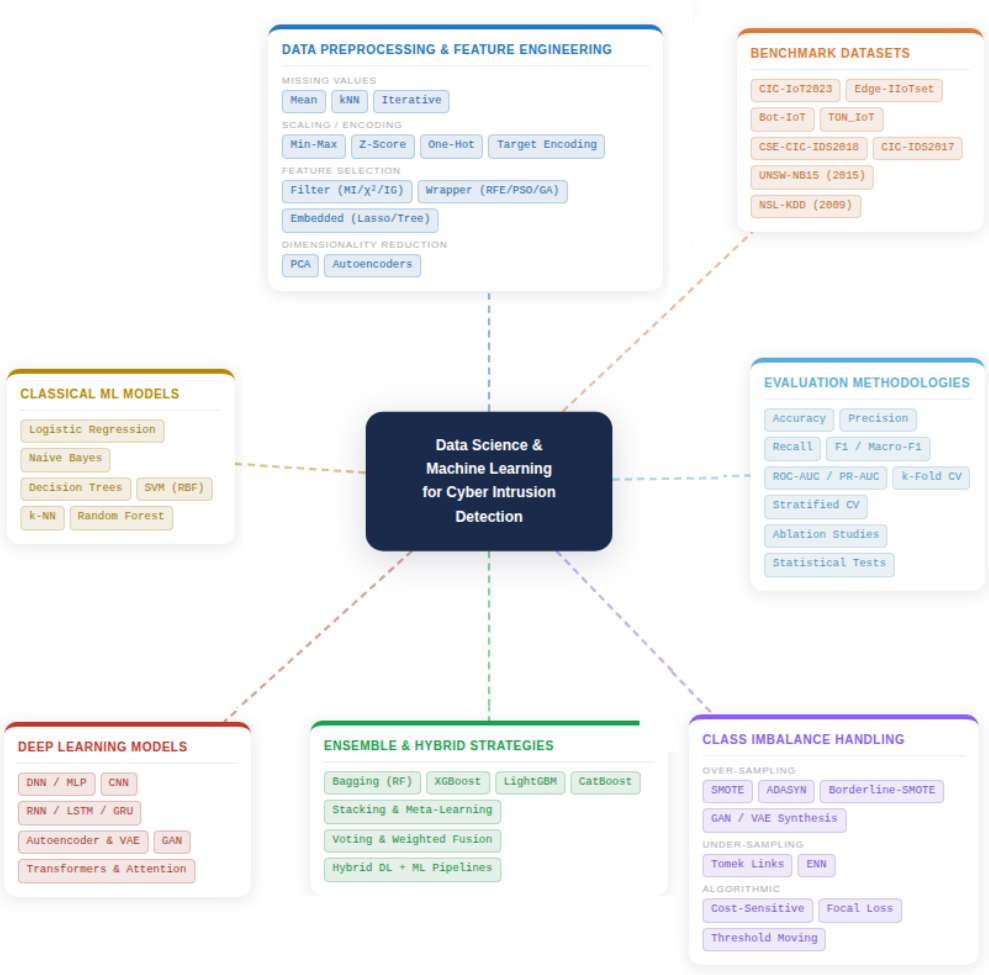

Figure 1. *Comprehensive Concept Map - Seven Major Data Science & Machine Learning Areas in Intrusion Detection.*

2. Background and Unified Taxonomy

2.1. Intrusion Detection Paradigms

Conventional IDS are categorized into signature-based, anomaly-based, and hybrid paradigms [5, 14]. Signature-based method computes the signature of existing data files and compares them against known attack signature pools. It cannot detect newly emerged or zero-day threats [5]. Anomaly-based IDS creates statistical profiles of normal traffic using entropy measures, clustering-based methods (e.g., isolation forests, Density-Based Spatial Clustering of Applications with Noise (DBSCAN)), and spectral measures to detect cyberattacks [15]. Anomaly-based methods offer the capability to detect newly appeared or zero-day attacks while having elevated false positive rates [15]. Hybrid paradigms integrate both engines, scanning high-confidence threat matches in the signature layer and forwarding other residual data traffic to the anomaly-based detecting system [5].

Recently, behavioral and contextual patterns detection has emerged beyond packet-level analysis for intrusion detection. Deep packet inspection is used to analyze the application-layer patterns and semantics, and flow-level features expose the patterns in the transport-layer without payload decryption [4]. Moreover, graph-based representations can be applied to model the host interaction patterns, which empowers the detection of lateral movement and slowly progressing APT campaigns [16]. Deception-based method also enables high-fidelity threat intelligence for intrusion detection with limitations in scalability and coverage [17].

2.2. Benchmark Datasets

The availability and quality of benchmark datasets are fundamental to IDS research. Table 1 summarizes the number of records, attack types, and key characteristics of the major datasets from 2009 to the present.

Table 1. *Summary of Major Benchmark Datasets for Network Intrusion Detection (2009–The Present).*

Dataset	Year	Records	Attack Types	Key Characteristics
NSL-KDD [18]	2009	148K	4 categories	Improved KDD'99; no redundant records; widely used legacy benchmark
UNSW-NB15 [9]	2015	2.54M	9 attack types	Modern attack types; 49 features; realistic network configuration
CIC-IDS2017 [10]	2017	2.83M	7 attack types	Full network traffic capture; labelled benign and attack flows
CSE-CIC-IDS2018 [19]	2018	16M+	7 attack types	Large scale; multi-day capture; AWS-based infrastructure
Bot-IoT [20]	2019	72M+	5 attack types	IoT-focused; botnet traffic; realistic smart environment
TON_IoT [21]	2020	461K	9 attack types	Heterogeneous IoT/IIoT telemetry; network and system data
CIC-IoT2023 [22]	2023	46M+	33 attack types	Latest IoT benchmark; 33 specific attack scenarios
Edge-IIoTset [23]	2022	20M+	14 attack types	Edge and IIoT environments; diverse protocols
CICIoMT2024 [24]	2024	Large	18 attack types	Internet of Medical Things; 40 devices across Wi-Fi, MQTT, and Bluetooth; DDoS, DoS, spoofing

The NSL-KDD dataset [18], despite its age, remains a widely used baseline whose feature distributions can not reflect modern traffic. UNSW-NB15 [9] incorporates modern attack types across 49 features, while CIC-IDS2017 [10] provides full network traffic captures. Recent datasets such as CIC-IoT2023 [22] and Edge-IIoTset [23] target IoT and edge computing environments, reflecting the evolving modern threat landscape. CICIoMT2024 [24] extends this trajectory to the Internet of Medical Things, capturing traffic from 40 devices across multiple protocols subjected to 18 attack types.

A persistent concern is that most benchmark datasets are generated in controlled laboratory environments, potentially not reflecting production-network statistical properties, adversarial evasion patterns, or annotation noise of the real-world networks [25, 26]. Recent survey work confirms that many studies rely on datasets with obsolete attack profiles [27, 28]. Dataset bias, feature leakage, and the absence of standardized evaluation protocols remain significant challenges [29].

2.3. A Unified Taxonomy for DS and ML-based IDS

To organize the heterogeneous IDS literature, the author proposes a unified four-axis taxonomy classifying each study along four orthogonal dimensions shown in Table 2. This four-axis taxonomy include the coverage of learning strategy, imbalance handling, explainability level, and deployment context.

Axis 1 – Learning Strategy. *Supervised* (labeled traffic classification), *Unsupervised* (clustering, autoencoders, one-class Support Vector Machine (SVM)), *Semi-supervised* (small labeled sets with large unlabeled corpora), *Self-supervised* (contrastive or generative pre-training), and *Reinforcement Learning* (RL, adaptive detection policies).

Axis 2 – Imbalance Handling. *None/implicit*, *Sampling-based* (Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), undersampling), *Cost-sensitive* (modified loss, focal loss, threshold-moving), *Ensemble-based* (recall-weighted voting, boosting), and *Generative* (GAN, Variational Autoencoder (VAE) for the synthetic minority augmentation).

Axis 3 – Explainability Level. *Intrinsic* (decision trees, rule-

based), *Post-hoc local* (SHAP, Local Interpretable Model-Agnostic Explanations (LIME)), *Post-hoc global* (feature importance rankings), and *Hybrid explainable ensemble* (attention-weighted stacking).

Axis 4 – Deployment Context. *Cloud/Datacenter*, *Edge/IoT* (resource-constrained), *Federated* (distributed privacy-preserving), *Software-Defined Networking (SDN)*, and *Industrial Cyber-Physical System (CPS) / Supervisory Control and Data Acquisition (SCADA)*.

Table 2. *Unified Four-Axis Taxonomy for Data Science and Machine Learning based Intrusion Detection Systems.*

Axis	Categories	Representative Works	Key Trade-offs
Learning Strategy	Supervised, Unsupervised, Semi-supervised, Self-supervised, RL	[30–37] (Supervised); [38] (Generative); [39] (Self-supervised)	Supervised: high accuracy, label-dependent. Unsupervised: zero-day capable, high FPR. Self-supervised: scalable, immature for IDS
Imbalance Handling	None/Implicit, Sampling-based, Cost-sensitive, Ensemble-based, Generative	[37] (SMOTE); [40] (Focal loss); [34, 36] (Ensemble recall)	Sampling: simple, may over-fit to synthetic minority. Cost-sensitive: no data modification and requires tuning. Generative: realistic but complex
Explainability Level	Intrinsic, Post-hoc local, Post-hoc global, Hybrid explainable ensemble	[41–43] (SHapley Additive exPlanations (SHAP)/LIME); [31] (Intrinsic); [44] (Retrieval-Augmented Generation (RAG)-XAI); [45] (Feature importance)	Intrinsic: interpretable but limited capacity. Post-hoc: any model, approximate only. Hybrid: best analyst utility, highest engineering cost
Deployment Context	Cloud/DC, Edge/IoT, Federated, SDN, Industrial CPS	[33, 35] (Cloud); [20–24] (IoT/Edge); [43, 46–49] (Federated)	Cloud: unlimited compute, centralization risk. Edge: constrained hardware. Federated: privacy-preserving, not independent and identically distributed (non-IID) challenges

3. Data Preprocessing and Feature Engineering

3.1. Data Preprocessing Pipelines

Data science preprocessing pipelines that precede model training play a critical role in the intrusion detection performance while being underreported in intrusion detection literature [50]. Raw network traffic data can have significant quality issues and typically contains redundant, noisy, and incomplete information [50]. Those issues include, but are not limited to duplicate records from flow aggregation, missing values from packet loss, feature scales covering orders of magnitude, and infinite values from rate computation division-by-zero [50, 51].

Missing values can be handled by imputation strategies, including mean/median substitution, K-Nearest Neighbor (KNN) imputer, and iterative imputation approaches [5]. Feature scaling methods, such as min-max normalization or z-score standardization, are applied to handle the issue of orders of magnitude feature value, which could be essential for distance-based learning methods (KNN, SVM) [52]. Tree-based methods are effective and invariant to the monotonic transformations. Moreover, feature encoding methods (one-hot, ordinal, and target encoding) can be used to encode semantic relationships between categorical values in the cyber intrusion dataset [14]. Meanwhile, the absence of standardized preprocessing methodologies makes it difficult to attribute observed performance differences to the pipeline or to the model architecture [50].

3.2. Feature Selection and Dimensionality Reduction

Feature selection is critical for reducing dimensionality, removing redundant attributes, lowering overfitting, and improving model generalization. Filter methods rank features via information gain, mutual information, and chi-squared test. They are independent of the learning algorithms and are important for intrusion detection [13]. Wrapper methods, including Recursive Feature Elimination (RFE) and genetic algorithms can be employed to optimize the feature subsets and model validation performance with possibly increased computational cost [53]. Embedded methods, including L1 regularization and tree-based feature importance, are effective options in incorporating feature selection within training methods to enhance model performance [45]. Principal Component Analysis (PCA) provides linear dimensionality reduction to remove redundant features. Autoencoders and Independent Component Analysis (ICA) provide the nonlinear alternatives for learning compact traffic representations [38].

Sarhan et al. [29] proposed a standard feature set selection for network IDS datasets, addressing cross-dataset heterogeneity that complicates transfer learning. Pigeon-inspired feature selection method can have potential in reducing the feature dimensionality while keeping detection accuracy [53]. Selvam and Latha [13] presented a PSO-GA hybrid optimization approach whose fitness function can explicitly penalize false negatives on the minority classes.

3.3. Embedding-Based and Contrastive Feature Learning

It has been an emerging paradigm that embedding-based and contrastive feature learning has been utilized for

intrusion detection [54]. Embedding-based representations and self-supervised contrastive learning can be leveraged to derive discriminative traffic features without extensive manual engineering [39]. Word2Vec-inspired approaches encode protocol sequences into continuous embedding spaces, while transformer-based encoders pre-trained on large-scale unlabeled traffic corpora enable strong transfer capability with fewer labeled samples [39]. Contrastive learning frameworks such as SimCLR and MoCo, which are adapted for IDS, train encoders to bring representations of similar flows closer and separate dissimilar ones [27]. This represents a fundamental shift toward jointly optimized representation learning for efficient intrusion detection.

4. Classical Machine Learning Methods

Machine learning has gained popularity on cyberattack detection and spans a broad range of algorithmic models. Naïve Bayes, logistic regression, and linear discriminant analysis have delivered computationally efficient and interpretable baselines for intrusion detection [7, 14]. Meanwhile, they encounter similar issues of nonlinear decision boundaries in the heterogeneous threat traffic [30]. K-Nearest Neighbor predicts the attack labels based on the local similarity with competitive performance on balanced datasets while suffering from a curse of dimensionality on high-dimensional cyberattack dataset [2]. Decision trees provide transparent and interpretable models with vulnerability to overfitting on complicated feature sets [30]. SVMs provide strong performance in model generalization by maximizing the margin and transforming features in the high-dimensional intrusion feature space [30, 32]. Gu and Lu [31] proposed combining SVM with Naïve Bayes feature embedding, achieving improved performance by leveraging probabilistic feature representations as input to the margin-based classifier.

Kilincer et al. [30] provided an extensive comparative study of conventional machine learning methods across multiple cyberattack datasets, evaluating the performance of decision tree, random forest, SVM, KNN, and Naïve Bayes models on intrusion detection. Comparative analysis has highlighted the dataset-dependent nature of model rankings, cautioning against universal claims of algorithmic superiority [26, 30]. Mishra et al.[15] found that the data preprocessing quality and feature selection strategy often produces greater influences on the final performance than the choice of classification algorithm, underscoring the critical importance of the data science pipeline for most intrusion detection research.

5. Deep Learning Approaches

In recent years, Deep learning algorithms and architectures have attracted substantial attention from many researchers in the intrusion detection field. Al-Garadi et al. [2] documented growing adoption of deep architectures across network security, malware detection, and threat intelligence applications. Multi-layer deep neural networks can achieve competitive accuracy on UNSW-NB15 and KDD Cup 1999, noting challenges with minority class detection [33]. While deep models deliver high benchmark accuracy, their computational overhead, black-box nature, and opacity still pose significant challenges for operational deployment [6].

Furthermore, Convolutional Neural Networks (CNNs) have been applied in intrusion detection, which process raw network traffic features as the spatial patterns and learn the correlations between local features [11]. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks model temporal dependencies in the sequential data flow, particularly relevant for detecting multi-step attack patterns. Choudhary and Kesswani [12] analyzed LSTM performance across the benchmark dataset KDD-Cup'99, NSL-KDD, and UNSW-NB15, reporting strong detection rates and noted sensitivity to hyperparameter tuning. Recent surveys reveal that hybrid CNN–LSTM architectures consistently achieve the best accuracy-efficiency trade-off across diverse IoT benchmarks dataset [28]. Autoencoder-based deep learning approaches can detect anomalies based on reconstruction error, with variational autoencoders extending this by imposing distributional constraints on the latent space, which can potentially improve generalization to unseen attack types [38].

6. Ensemble and Hybrid Strategies

Ensemble methods have been evolving as one of the most efficient paradigm for intrusion detection [45]. Random Forest (RF) trains multiple decision trees on bootstrap instances exhibiting high robustness to overfitting and noisy features [34]. Gradient boost methods, such as AdaBoost and XGBoost, can iteratively fit shallow trees to residuals and achieve the cutting-edge performance on the benchmark datasets [35]. Dhaliwal et al. [35] presented XGBoost has advantages over neural network and SVM baselines, which can be attributed to its superiority in regularization and distinct feature selection mechanisms.

Stacking method utilizes a meta-learner that is trained on out-of-fold predictions from different base models, enabling it to learn complementary error patterns [36]. Rajagopal et al. [36] demonstrated the effectiveness and efficiency of heterogeneous stacking strategy across multiple intrusion detection datasets. Tama et al. [32] presented a two-stage classifier (TSE-IDS) that identifies anomalous records first and then categorizes them by different attack types, which achieved strong accuracy performance on the dataset UNSW-NB15.

Standard majority voting or simple averaging are prone to favor and prioritize majority-class predictions, resulting in lower detection in minority classes [34]. Recall-aware weighting schemes can explicitly balance sensitivity across imbalanced classes by assigning higher ensemble weights to models, demonstrating superior recall on minority but critical attack categories, such as Worms and Shellcode [34, 36]. Stage-gated data processing pipeline that produces validated

outputs from each preprocessing phase can further amplify the benefits of recall-aware weighting by ensuring clean inputs [45, 50].

7. Quantitative Meta-Analysis and Class Imbalance in IDS

7.1. Quantitative Meta-Analysis of the IDS Literature

The author performed a quantitative meta-analysis of the 153 reviewed papers published from 2009 to 2025 along four dimensions: primary dataset, primary algorithm family, imbalance handling strategy, and evaluation metric. Results are summarized in Table 3. UNSW-NB15 and CIC-IDS2017 jointly dominate at 71% combined frequency, reflecting the community's reliance on two specific benchmarks. NSL-KDD still appears in 34% of papers, often as a secondary validation set despite known limitations [18]. IoT-specific datasets appear in 28% of papers, a proportion that has grown sharply since 2022. Real-world or proprietary traffic is used in fewer than 8% of studies, confirming the benchmark monoculture problem.

Table 3. *Dataset Frequency and Model-Type Distribution Across 153 Reviewed IDS Studies.*

Dataset Usage Frequency			Algorithm Family Distribution		
Dataset	**Papers**	**% of corpus**	**Algorithm Family**	**Papers**	**% of corpus**
UNSW-NB15	71	46%	Ensemble (RF/XGB/Stack)	54	35%
CIC-IDS2017/2018	38	25%	Deep Learning (CNN/LSTM/AE)	61	40%
NSL-KDD	52	34%	Classical ML (SVM/DT/NB)	28	18%
Bot-IoT / TON_IoT	31	20%	Transformer / Attention	12	8%
CIC-IoT2023 / Edge-IIoTset	12	8%	GNN-Based	7	5%
KDD Cup 99	41	27%	LLM / RAG-Based	4	3%
Real-world / Proprietary	12	8%	Federated Learning	9	6%
CICIoMT2024	3	2%	Hybrid Multi-strategy	31	20%

Deep learning families represent the single largest algorithmic category at approximately 40% of the corpus, while ensemble methods account for 35% [35, 36]. Classical ML approaches appear in 18% of papers, typically as baselines. Transformer and attention-based models constitute only 8% despite their theoretical fit for sequential traffic data. GNN-based and LLM/RAG-based IDS account for 5% and 3% respectively, indicating that the most architecturally advanced paradigms remain nascent. Hybrid multi-strategy approaches appear in 20% of papers. The proportion of papers employing explicit imbalance handling is 62%, but only 21% explicitly optimize for per-class recall rather than aggregate accuracy, motivating the dedicated analysis in Section 9. The most common evaluation metrics are accuracy (89% of papers), F1-score (71%), and Area Under the Receiver Operating Characteristic Curve (AUC-ROC, 58%), while per-class recall for minority attack types is reported in only 34% of papers.

7.2. Handling Class Imbalance in IDS

Class imbalance constitutes a pervasive and poses severe challenges in the intrusion detection datasets. Minority attack categories, such as Worms and Backdoors, can account for an extremely small percentage of the total network traffic [9, 32]. Standard learners can be biased toward favoring the majority attack class, gaining high overall detection accuracy while misclassifying a significant portion of the minority-class cyberattacks [5].

7.2.1. The Minority Attack Collapse Phenomenon

A fundamental asymmetry exists concerning the potential cost of errors in intrusion classification. A missed attack (false negative) can result in data exfiltration, security breach, loss of productivity, and even litigation [55]. In contrast, a wrongly detected attack (false positive) can possibly increase Security Operations Center (SOC) workload while rarely causing direct harm or damage. This asymmetry mandates *recall-centric* evaluation while not gaining strong research attention, and only 34% of reviewed papers report per-class recall [5].

The *minority attack collapse* phenomenon occurs when an imbalanced classifier achieves an apparently high overall accuracy by abstaining from effectively predicting the minority class [5]. If Worm attacks constitute 0.1% of traffic, a classifier can label all flows as benign with 99.9% accuracy while providing zero security value. Tama et al. [32] documented detection methods where models that achieved 98% aggregate accuracy yielded fewer than 40% recall on Shellcode and Backdoor categories. Therefore, a rigorous evaluation framework should report per-class precision and recall, confusion matrices, and Macro-averaged F1, specifying minimum acceptable recall thresholds per attack class aligned with the operational threat model.

7.2.2. Data-Level and Algorithm-Level Approaches

Synthetic Minority Over-sampling Technique (SMOTE) and its variant method, including ADASYN, Borderline-SMOTE, SMOTE-ENN, remain the widely used oversampling approaches to handle class imbalance and minority attack collapse issue [37]. Undersampling methods can reduce

majority class representation through random removal or Tomek links. Cost-sensitive learning modifies the loss function and assigns higher misclassification costs to minority classes for balancing the class structure and learning objectives [32]. Focal loss, adapted for IDS, reduces the weights of well-classified majority examples and concentrates learning on hard minority cases [40]. Threshold-moving can adjust decision boundaries post-training to optimize recall-precision trade-offs [30].

7.2.3. Generative and Self-Supervised Approaches

GANs and variational autoencoders synthesize realistic minority-class samples, capturing complex feature distributions more faithfully than SMOTE's linear interpolation [38]. Self-supervised contrastive pre-training enables models to cluster semantically similar traffic flows while separating dissimilar ones in the representation space, which yield features more robust to distribution shifts introduced by novel attack variants [27]. These approaches are valuable for online intrusion detection where models should adapt to continuously evolving attack traffic without requiring full retraining. A neglected dimension of imbalance handling is the economic cost model in which different attack categories impose fundamentally various operational costs [45]. A principled imbalance strategy should assign per-attack-class false-negative costs aligned with organizational threat models, rendering cost parameters as configurable SOC operator inputs [45, 50].

8. Challenges and Research Gaps

Based on the comprehensive literature review and quantitative meta-analysis, the author identifies nine technically grounded research gaps that are summarized in Table 4.

Table 4. *Technical Open Challenges and Research Gaps in Data Science and Machine Learning based Cyber Intrusion Detection.*

No.	Research Gap	Relevant Area	Key Issue
G1	Lack of standardized preprocessing and ablation methodology	Data Science & Methodology	Preprocessing is underreported; component-level contributions are rarely isolated via ablation
G2	Insufficient real-world and cross-dataset evaluation	Datasets & Evaluation	Most studies use single synthetic benchmarks; real-world generalization is unverified
G3	Inadequate minority-class recall optimization	Class Imbalance	Most studies optimize overall accuracy; rare attack types remain poorly detected
G4	Limited adversarial robustness evaluation	Security	Few IDS studies evaluate robustness against adversarial evasion attacks
G5	Absence of online, streaming, and edge-deployable frameworks	Deployment	Batch models lack concept-drift adaptation; IoT/edge constraints are under-addressed
G6	Insufficient explainability for SOC operations	XAI	Complex models lack actionable, analyst-oriented explanations
G7	Underdeveloped federated and privacy-preserving IDS	Privacy	Collaborative cross-organization training without data sharing remains immature
G8	Nascent application of transformers, LLMs, and GNNs	Emerging Models	Advanced architectures for structured flow-data IDS are still early-stage
G9	Limited zero-shot, few-shot, and transfer learning for IDS	Generalization	Rapid adaptation to novel attack types with scarce labels is under-explored

G1: Standardized Preprocessing. Data preprocessing remains one of the most underreported elements in IDS research [50]. Many studies describe data preprocessing incompletely, making reproduction impossible. Evidence suggests that data preprocessing improvements can explain over 50% of the performance gap between baseline and state-of-the-art results. The resolution should establish open-source and stage-gated preprocessing frameworks and mandate ablation experiments in multi-component pipelines [51].

G2: Real-World and Cross-Dataset Evaluation. Most IDS studies fail to evaluate on production dataset with concept drift, encrypted payloads, or annotation noise absent from laboratory settings [25, 26]. Few studies perform cross-dataset validation, and the lack of standardized feature sets complicates the assessment of generability and transferability [29]. Bridging this gap requires industry and academia partnerships for privacy-preserving production-traffic evaluation.

G3: Minority-Class Recall. Although SMOTE and its variants are commonly applied, a majority studies optimize aggregate accuracy or macro-F1 instead of targeting per-class recall [37]. Rare but dangerous attack types remain poorly detected even when aggregate metrics appear strong. One promising work is to develop recall-aware ensemble optimization frameworks that expose per-class detection thresholds as configurable operator inputs [34].

G4: Adversarial Robustness. Despite known machine learning vulnerability to adversarial examples, very few studies include robustness evaluation as a performance metric [56]. Current research focuses predominantly on white-box attacks, while query-based and transfer-based black-box threat models remain underexplored [57]. Certified robustness guarantees and domain-adapted adversarial training require

further investigation.

G5: Online, Streaming, and Edge Deployment. Most data science and machine learning based IDS operate in batch mode, but real-world traffic is non-stationary with continuous concept drift [25]. Drift detection methods (e.g., Adaptive Windowing (ADWIN), Page-Hinkley) have been used in intrusion detection, but their integration into end-to-end pipelines with automated retraining triggers is underdeveloped. Lightweight model compression (e.g., pruning, quantization, knowledge distillation) for IoT edge deployment remains an active challenge [20–23].

G6: Explainability for SOC Operations. Explainable Artificial Intelligence (XAI) techniques, such as SHAP and LIME, provide feature-importance scores, but translating these into actionable domain-specific guidance for SOC analysts remains a challenging issue [41, 42]. Arreche et al. [43] demonstrated that integrating SHAP explanations with XGBoost improves analyst trust while maintaining detection accuracy. However, its impact on analyst response time and false-positive triage rates remains unmeasured.

G7: Federated and Privacy-Preserving IDS. Federated learning enables collaborative IDS training without sharing raw data while facing non-IID convergence difficulties, communication constraints, and vulnerability to poisoning attacks [46]. Communication constraints limit update frequency and the vulnerability to adversarial participants who inject poisoned updates poses security risks. Differential privacy, while offering formal guarantees, may degrade rare-attack detection. The pFedCross framework [49] and FD-IDS [48] have begun addressing these challenges, but most works use artificially balanced splits that contradict realistic non-IID federated scenarios [47].

G8: Transformers, LLMs, and GNNs. Transformer self-attention has applicability for processing sequential flow data while posing quadratic complexity at high throughput [39]. Pre-training strategies for network traffic domains remain immature, and LLM/RAG application to real-time flow classification is largely unexplored [58]. GNNs model can host interaction topology for detecting lateral movement, but its scalable dynamic-graph architectures for streaming traffic are underdeveloped. Significant open challenges include hallucination risk, knowledge base poisoning, and inference latency incompatible with real-time requirements [27, 39].

G9: Zero-Shot, Few-Shot, and Transfer Learning. Novel attack types with minimal labeled data demand rapid adaptation capabilities. Transfer learning from labeled benchmarks to new target networks, metric-based few-shot learning, and meta-learning for rapid attack-type recognition remain under-explored in IDS [59]. These approaches are critical for zero-day detection [60] and cold-start deployment scenarios.

9. Emerging Trends and An Actionable Architecture

This section surveys eight emerging trends reshaping data science and machine learning based intrusion detection. Table 5 provides a concise summary of each paradigm, its core advantage, and current limitations.

Table 5. *Summary of Emerging Trends and New Paradigms in Data Science and Machine Learning based Intrusion Detection.*

Paradigm	Core Advantage	Key Techniques	Open Challenges
Transformers & Attention	Captures long-range temporal dependencies	Self-attention, hybrid CNN/LSTM–transformer models	Quadratic complexity; immature pre-training for traffic data
LLMs & Foundation Models [58]	Contextual reasoning over logs, alerts, and threat intelligence	BERT-based classifiers, RAG pipelines	High latency and cost; unproven for real-time flow classification
Federated Learning [46]	Privacy-preserving collaborative training	FedAvg, differential privacy, secure aggregation	Non-IID distributions; communication cost; poisoning attacks
Adversarial Robustness [56]	Ensures IDS resilience against evasion attacks	Adversarial training, randomized smoothing, ensemble diversity	Domain constraint of preserving attack functionality; limited black-box evaluation
Explainable AI [41, 42]	Builds SOC analyst trust via interpretable detections	SHAP, LIME, attention visualization, surrogate models	Translating feature scores into actionable analyst guidance
Zero/Few-Shot & Transfer Learning	Rapid adaptation to novel attacks with scarce labels	Meta-learning, metric learning, domain adaptation	Under-explored in IDS; lack of episodic training protocols
IoT & Edge IDS [20–23]	Detection under severe resource constraints	Lightweight models, pruning, edge–cloud hierarchies	Heterogeneous protocols; concept drift at scale; sector-specific gaps
Agentic LLM-RAG IDS [44, 61]	Real-time attack classification and SOC alert enrichment	Agentic RAG pipelines, fine-tuned classifiers, knowledge bases	Hallucination risk; knowledge base poisoning; inference latency

9.1. Emerging Tends and New Paradigms

Transformers and Attention. Transformer self-attention captures long-range dependencies in sequential flow data without vanishing gradient limitations of RNNs [62]. Both pure transformer and hybrid attention–CNN/LSTM architectures show promise for detecting multi-step attacks [39].

LLMs and Foundation Models. LLMs have been explored for threat intelligence analysis, log parsing, and alert triage, with Retrieval-Augmented Generation (RAG) architectures combining LLM reasoning with threat databases [58]. Houssel et al. [44] demonstrated that LLMs can provide meaningful explanations for NetFlow-based detection through RAG, bridging the interpretability gap between ML detections and SOC analyst comprehension. The CyberRAG framework [61] advances this with a fully agentic architecture delivering real-time attack classification, structured incident reports, and contextual mitigation recommendations, addressing alert fatigue in large enterprise environments.

Graph Neural Networks (GNNs). GNNs model traffic as graphs, representing hosts as nodes and communications as edges. That enables detection of relational intrusion patterns such as lateral movement and coordinated scanning [63, 64]. Graph-based community detection and flow correlation can reveal attacks that are invisible to a flat feature-vector analysis [63].

Federated and Privacy-Preserving Learning. Federated learning enables collaborative IDS training without sharing raw traffic data [46]. The pFedCross framework [49] tackles non-IID heterogeneity through cross-organization model sharing with Jensen-Shannon divergence-based distribution optimization. FD-IDS [48] combines knowledge distillation with proximal regularization for non-IID IoT environments. A SHAP-based federated framework preserves raw data privacy while enabling feature-importance sharing, simultaneously addressing federated convergence and explainability [27].

Adversarial Robustness. Adversarial perturbations to flow features can evade machine learning and deep learning detectors while preserving attack functionality [56]. Defense strategies include adversarial training, randomized smoothing, and ensemble diversity, though realistic black-box threat models remain under-explored [57].

Explainable AI (XAI) for Intrusion Detection. XAI techniques such as SHAP, LIME, and attention visualization provide feature-level and instance-level detection explanations [41, 42]. The key gap is translating technical importance scores into actionable, domain-specific guidance for SOC analysts [41].

Zero-Shot, Few-Shot, and Transfer Learning. Transfer learning, metric-based few-shot learning, and meta-learning address novel attack types with scarce labels [59]. These approaches are critical for zero-day detection and cold-start deployment while remaining under-explored in the IDS context[60].

IoT and Edge-Based IDS. IoT environments impose severe resource constraints demanding lightweight models and hierarchical edge–cloud architectures [20–24]. CICIoMT2024 [24] provides a dedicated benchmark for the Internet of Medical Things with 18 attack types, enabling IDS solutions tailored to clinical environment requirements. Model compression techniques, including structured pruning, quantization-aware training, and knowledge distillation, are essential enablers for deploying accurate IDS models on microcontroller-class IoT hardware [28].

9.2. Cutting-Edge IDS: An Actionable Architecture

Drawing on the synthesis of this review, the author proposes a seven-stage actionable architecture for the state-of-the-art data science and machine learning based intrusion detection system shown in Figure 2. This architecture represents a forward-looking integration of the cutting-edge technology advances identified in our analysis.

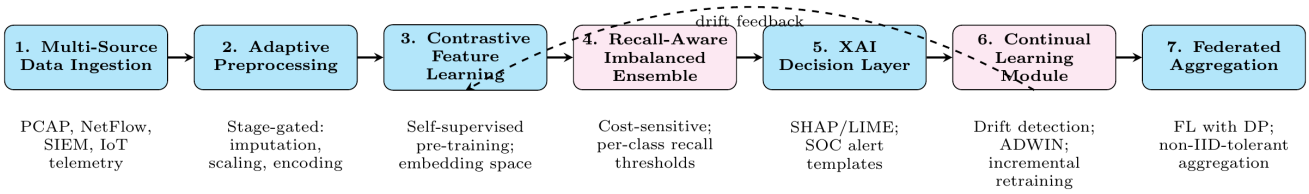

Figure 2. *Actionable seven-stage architecture for next-generation data science and machine learning based IDS. Blue stages represent data and feature processing; pink stages represent detection and adaptation components. The dashed feedback arrow from Stage 6 (Continual Learning) to Stage 3 (Feature Learning) enables representation refresh upon concept drift.*

Stage 1: Multi-Source Data Ingestion. Accepts heterogeneous inputs including PCAP, NetFlow/IPFIX records, Security Information and Event Management (SIEM) log streams, and IoT device telemetry, normalizing protocol-specific representations into a common event schema.

Stage 2: Adaptive Data Science Preprocessing Pipeline. A stage-gated layer applies sequential quality validation checkpoints: missing value detection and imputation, infinite-value handling, duplicate removal, feature scaling, and categorical encoding. Data flows failing quality thresholds are quarantined instead of being silently propagated (addressing G1).

Stage 3: Contrastive Feature Learning. A self-supervised contrastive encoder pre-trained on unlabeled network traffic learns a semantically rich embedding space where benign and attack data flows are well-separated without requiring labels at pre-training time. Transfer learning enables rapid adaptation to new environments (addressing G9).

Stage 4: Recall-Aware Imbalanced Ensemble. The detection core is an ensemble model with configurable operator per-class cost weights. A meta-learning layer weights component classifiers based on per-class recall on a held-out validation set, automatically prioritizing detectors with high recall on rare attack categories (addressing G3).

Stage 5: Explainable AI Decision Layer. Every detection decision is accompanied by a SHAP/LIME based local explanation mapped to a domain-specific SOC alert template. Global feature importance rankings and concept-level summaries should also be provided, empowering the trust in calibration process from SOC analysts (addressing G6).

Stage 6: Continual Learning Module. An online drift detector (e.g., ADWIN, Page-Hinkley) monitors the feature distribution of incoming traffic. Upon detecting significant concept drift, an incremental retraining pipeline updates the contrastive encoder (Stage 3) and ensemble weights (Stage 4) using a sliding replay buffer without full retraining (addressing G5).

Stage 7: Federated Aggregation. In multi-organization deployment scenarios, a privacy-preserving federated aggregation server collects model updates using differential privacy (DP) and non-IID-tolerant algorithms, which can combine local knowledge without centralizing sensitive traffic data (addressing G7).

10. Conclusion and Recommendation

This paper presented a systematic review of data science and machine learning applications in cyber intrusion detection, covering 153 studies from 2009 to 2025. A unified four-axis taxonomy classified IDS research by learning strategy, imbalance handling, explainability, and deployment context. The review surveyed data preprocessing and feature engineering techniques, classical machine learning models, deep learning architectures, ensemble and hybrid strategies, class imbalance handling methods, benchmark datasets, open challenges, and emerging paradigms.

The author identified nine technically grounded research gaps spanning preprocessing standardization, cross-dataset evaluation, minority-class recall, adversarial robustness, online and edge deployment, explainability for SOC operations, federated learning, transformer/LLM/GNN application, and zero-shot adaptation. In response, the author proposed a seven-stage actionable and state-of-the-art IDS architecture integrating adaptive preprocessing, contrastive representation learning, recall-aware imbalanced ensemble detection, XAI decision layer, continual learning with drift detection, and privacy-preserving federated aggregation. Each stage directly addresses identified gaps, providing a concrete roadmap from current benchmark-focused research to operational real-world deployment.

Future research directions of particular priority include the following. First, the development of open-source, standardized preprocessing frameworks with quality checkpoints that enable reproducible and fair comparison across studies [50]. Second, the creation of realistic evaluation environments and privacy-preserving protocols that bridge the gap between benchmark performance and operational effectiveness. Third, the investigation of transformer architectures and large language models for real-time traffic classification, potentially integrating Retrieval-Augmented Generation (RAG) for context-aware detection [44, 61]. Fourth, the advancement of federated learning approaches that handle non-IID traffic distributions and adversarial participants while maintaining differential privacy guarantees [47, 49]. Fifth, the establishment of ablation study standards that require component-level contribution analysis in multi-stage IDS pipelines. Addressing these gaps will require interdisciplinary collaboration between machine learning researchers, network security practitioners, and SOC analysts, bringing together theoretical advances with operational requirements in intrusion detection.

Additionally, the author recommends that future studies should incorporate the exploration of intensive accuracy benchmarking of the ML and DL models reviewed across standardized datasets with consistent preprocessing protocols. A comprehensive comparative accuracy analysis of classical ML models, deep learning architectures, ensemble strategies, and emerging paradigms (transformers, GNNs, LLMs) under identical experimental conditions is also valuable future work. An accuracy prediction for the existing frameworks and proposed architecture would provide significant technical value and establish reliable performance baselines for researchers and developers to build robust and reliable intrusion detection systems.

Abbreviations

ADASYN	Adaptive Synthetic Sampling
ADWIN	Adaptive Windowing
APTs	Advanced Persistent Threats
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CNNs	Convolutional Neural Networks
CPS	Cyber-Physical System
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DL	Deep Learning
DP	Differential Privacy
DS	Data Science
GAN	Generative Adversarial Network
GNN	Graph Neural Network
ICA	Independent Component Analysis
IDS	Intrusion Detection System
IoT	Internet of Things
KNN	K-Nearest Neighbor
LIME	Local Interpretable Model-Agnostic Explanations
LLMs	Large Language Models
LSTM	Long Short-Term Memory

ML	Machine Learning
non-IID	Non-Independent and Identically Distributed
PCA	Principal Component Analysis
PSO-GA	Particle Swarm Optimization–Genetic Algorithm
RAG	Retrieval-Augmented Generation
RFE	Recursive Feature Elimination
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SCADA	Supervisory Control and Data Acquisition
SDN	Software-Defined Networking
SHAP	SHapley Additive exPlanations
SIEM	Security Information and Event Management
SMOTE	Synthetic Minority Over-sampling Technique
SOC	Security Operations Center
SVM	Support Vector Machine
VAE	Variational Autoencoder
XAI	Explainable Artificial Intelligence

Author Contributions

Yali Ren: Conceptualization, Methodology, Investigation, Data Curation, Formal Analysis, Writing – Original Draft, Writing – Review & Editing, Visualization

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] S. Corbet and J. W. Goodell,"The reputational contagion effects of ransomware attacks," *Finance Research Letters*, vol. 47, pp. 102715, 2022. https://doi.org/10.1016/j.frl.2022.102715

[2] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine learning and deep learning in cybersecurity," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1646–1685, 2020. https://doi.org/10.1109/COMST.2020.2994934

[3] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity data science: An overview from machine learning perspective," *Journal of Big Data*, vol. 7, no. 1, p. 41, 2020. https://doi.org/10.1186/s40537-020-00318-5

[4] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of ML and DL approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, e4150, 2021. https://doi.org/10.1002/ett.4150

[5] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, p. 20, 2019. https://doi.org/10.1186/s42400-019-0038-7

[6] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study," *Journal of Information Security and Applications*, vol. 50, 102419, 2020. https://doi.org/10.1016/j.jisa.2019.102419

[7] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, p. 160, 2021. https://doi.org/10.1007/s42979-021-00592-x

[8] A. Thakkar and R. Lohiya, "A survey on intrusion detection system: Feature selection, model, performance measures, application perspective and challenges," *Artificial Intelligence Review*, vol. 55, pp. 453–563, 2022. https://doi.org/10.1007/s10462-021-10037-9

[9] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in *Proc. MilCIS*, pp. 1–6, 2015. [Dataset updated 2023.] https://doi.org/10.1109/MilCIS.2015.7348945

[10] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. ICISSP*, pp. 108–116, 2018. [CIC-IDS2017 dataset.] https://doi.org/10.5220/0006639801080116

[11] O. H. Abdulganiyu, T. A. Tchakoucht, and Y. K. Saheed, "A systematic literature review on network intrusion detection based on deep learning," *Expert Systems with Applications*, vol. 215, 119357, 2023. https://doi.org/10.1016/j.eswa.2022.119357

[12] S. Choudhary and N. Kesswani, "Analysis of KDD-Cup'99, NSL-KDD and UNSW-NB15 using deep learning with LSTM," *Procedia Computer Science*, vol. 167, pp. 592–600, 2020. https://doi.org/10.1016/j.procs.2020.03.068

[13] P. Selvam and K. Latha, "An efficient feature selection method for network intrusion detection using PSO-GA," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 1, pp. 1–15, 2023. https://doi.org/10.3233/JIFS-213095

[14] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," *Applied Sciences*, vol. 9, no. 20, 4396, 2019. https://doi.org/10.3390/app9204396

[15] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A detailed investigation and analysis of using machine learning techniques for intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 686–728, 2019. https://doi.org/10.1109/COMST.2018.2847722

[16] W. Villegas-Ch, J. Govea, A. Maldonado Navarro, and P. Palacios Játiva, "Intrusion Detection in IoT Networks Using Dynamic Graph Modeling and Graph-Based Neural Networks," IEEE Access, vol. 13, pp. 65356-65375, 2025. https://doi.org/10.1109/ACCESS.2025.3559325

[17] Z. Z. Shah, M. Ikram, H. J. Asghar, and M. A. Kaafar, "Deception Meets Diagnostics: Deception-based Real-Time Threat Detection in Healthcare Web Systems," in Proc. 28th International Symposium on Research in Attacks, Intrusions and Defenses (RAID), pp. 391-410, 2025. IEEE.

[18] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE CISDA*, pp. 1–6, 2009. https://doi.org/10.1109/CISDA.2009.5356528

[19] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy," in *Proc. IEEE ICCST*, pp. 1–8, 2019. https://doi.org/10.1109/ICCST.2019.8888419

[20] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the Internet of Things for network forensic analytics: Bot-IoT dataset," *Future Generation Computer Systems*, vol. 100, pp. 779–796, 2019. https://doi.org/10.1016/j.future.2019.04.026

[21] N. Moustafa, "A new distributed architecture for evaluating AI-based security systems at the edge: Network TON_IoT datasets," *Sustainable Cities and Society*, vol. 72, 102994, 2021. https://doi.org/10.1016/j.scs.2021.102994

[22] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment," *Sensors*, vol. 23, no. 13, 5941, 2023. https://doi.org/10.3390/s23135941

[23] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," *IEEE Access*, vol. 10, pp. 40281–40306, 2022. https://doi.org/10.1109/ACCESS.2022.3165809

[24] S. Dadkhah, E. C. P. Neto, R. Ferreira, R. C. Molokwu, S. Sadeghi, and A. A. Ghorbani, "CICIoMT2024: A benchmark dataset for multi-protocol security assessment in IoMT," *Internet of Things*, vol. 28, 101351, 2024. https://doi.org/10.1016/j.iot.2024.101351

[25] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE S&P*, pp. 305–316, 2010. [Reprinted with commentary 2019.] https://doi.org/10.1109/SP.2010.25

[26] G. Apruzzese, P. Laskov, M. de Oca Montes, W. Malber, and F. Roli, "The role of machine learning in cybersecurity," *Digital Threats: Research and Practice*, vol. 4, no. 1, pp. 1–38, 2023. https://doi.org/10.1145/3545574

[27] S. Hozouri, A. Rezaei, M. Bahrololoum, and S. M. Hosseinirad, "A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges," *Discover Artificial Intelligence*, vol. 5, p. 314, 2025. https://doi.org/10.1007/s44163-025-00578-1

[28] M. M. Rahman, S. A. Shakil, and M. R. Mustakim, "A survey on intrusion detection system in IoT networks," *Cyber Security and Applications*, vol. 3, 100082, 2025. https://doi.org/10.1016/j.csa.2024.100082

[29] M. Sarhan, S. Layeghy, and M. Portmann, "Towards a standard feature set for network intrusion detection system datasets," *Mobile Networks and Applications*, vol. 27, pp. 357–370, 2022. https://doi.org/10.1007/s11036-021-01843-0

[30] I. F. Kilincer, F. Ertam, and A. Sengur, "Machine learning methods for cyber security intrusion detection: Datasets and comparative study," *Computer Networks*, vol. 188, 107840, 2021. https://doi.org/10.1016/j.comnet.2021.107840

[31] J. Gu and S. Lu, "An effective intrusion detection approach using SVM with naïve Bayes feature embedding," *Computers & Security*, vol. 103, 102158, 2021. https://doi.org/10.1016/j.cose.2020.102158

[32] B. A. Tama, M. Comuzzi, and K. H. Rhee, "TSE-IDS: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system," *IEEE Access*, vol. 7, pp. 94497–94507, 2019. https://doi.org/10.1109/ACCESS.2019.2928048

[33] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525–41550, 2019. https://doi.org/10.1109/ACCESS.2019.2895334

[34] X. Gao, C. Shan, C. Hu, Z. Niu, and Z. Liu, "An adaptive ensemble machine learning model for intrusion detection," *IEEE Access*, vol. 7, pp. 82512–82521, 2019. https://doi.org/10.1109/ACCESS.2019.2924546

[35] S. S. Dhaliwal, A. A. Nahid, and R. Abbas, "Effective Intrusion Detection System Using XGBoost," Information, vol. 9, no. 7, 149, 2018. https://doi.org/10.3390/info9070149

[36] S. Rajagopal, K. S. Hareesha, and P. P. Mandya, "A stacking ensemble for network intrusion detection using heterogeneous datasets," *Security and Communication Networks*, 2020, 4586875. https://doi.org/10.1155/2020/4586875

[37] S. M. Kasongo and Y. Sun, "Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset," *Journal of Big Data*, vol. 7, no. 1, p. 105, 2020. https://doi.org/10.1186/s40537-020-00376-9

[38] T. Saba, A. Rehman, T. Sadad, H. Kolivand, and S. A. Bahaj, "Anomaly-based intrusion detection system for IoT networks through deep learning model," *Computers and Electrical Engineering*, vol. 99, 107810, 2022. https://doi.org/10.1016/j.compeleceng.2022.107810

[39] Z. Xu, Y. Wu, S. Wang, J. Gao, T. Qiu, Z. Wang, H. Wan, and X. Zhao, "Deep learning-based intrusion detection systems: A survey," arXiv: 2504.07839, 2025.

[40] M. Injadat, A. Moubayed, A. B. Nassif, and A. Shami, "Multi-stage optimized machine learning framework for network intrusion detection," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1803–1816, 2021. https://doi.org/10.1109/TNSM.2020.3014929

[41] A. Capuano, G. Chiarella, R. Gallo, and G. A. Angelini, "Explainable artificial intelligence in cybersecurity: A survey," *IEEE Access*, vol. 10, pp. 93575–93600, 2022. https://doi.org/10.1109/ACCESS.2022.3204171

[42] I. H. Sarker, "Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective," *SN Computer Science*, vol. 2, no. 3, p. 154, 2021. https://doi.org/10.1007/s42979-021-00594-9

[43] O. Arreche, T. Guntur, and M. Abdallah, "Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability," *Frontiers in Computer Science*, vol. 7, 1520741, 2025. https://doi.org/10.3389/fcomp.2025.1520741

[44] P. Houssel, P. Singh, S. Layeghy, and M. Portmann, "Towards explainable network intrusion detection using large language models," arXiv:2408.04342, 2024. https://doi.org/10.48550/arXiv.2408.04342

[45] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, "Building an efficient intrusion detection system based on feature selection and ensemble classifier," *Computer Networks*, vol. 174, 107247, 2020. https://doi.org/10.1016/j.comnet.2020.107247

[46] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021. https://doi.org/10.1016/j.future.2020.10.007

[47] A. Karim, R. Islam, M. Shahriar, M. S. Kaiser, S. A. Pirani, and S. Belhaouari, "Intrusion detection based on federated learning: A systematic review," *ACM Computing Surveys*, 2025. https://doi.org/10.1145/3731596

[48] Y. Djenouri, A. Belhadi, and P. Srivastava, "Federated learning-based intrusion detection in IoT networks: Performance evaluation and data scaling study," *Journal of Sensor and Actuator Networks*, vol. 14, no. 4, p. 78, 2025. https://doi.org/10.3390/jsan14040078

[49] X. Liu, T. Li, P. Gu, and Y. Su, "Survey of federated learning in intrusion detection," *Journal of Parallel and Distributed Computing*, vol. 195, 104977, 2025. https://doi.org/10.1016/j.jpdc.2024.104977

[50] R. S. M. L. Patibandla, S. S. Kurra, and N. B. Mundukur, "A study on scalability of data science pipeline for unstructured data," *Journal of Big Data*, vol. 9, no. 1, pp. 1–18, 2022. https://doi.org/10.1186/s40537-022-00564-9

[51] V. Kumar, A. K. Das, and D. Sinha, "Statistical analysis driven optimized neural network for intrusion detection," *International Journal of Information Security*, vol. 20, pp. 827–841, 2021. https://doi.org/10.1007/s10207-021-00544-7

[52] F. E. Ayo, S. O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, "Network intrusion detection based on deep learning model optimized with rule-based hybrid feature selection," *Inf. Secur. J.*, vol. 29, no. 6, pp. 267–283, 2020. https://doi.org/10.1080/19393555.2020.1767240

[53] H. Alazzam, A. Sharieh, and K. E. Sabri, "A feature selection algorithm for intrusion detection system based on pigeon inspired optimizer," *Expert Systems with Applications*, vol. 148, 113249, 2020. https://doi.org/10.1016/j.eswa.2020.113249

[54] Y. Yue, X. Chen, Z. Han, X. Zeng, and Y. Zhu, "Contrastive learning enhanced intrusion detection," *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 4232–4247, 2022. https://doi.org/10.1109/TNSM.2022.3218843

[55] Y. Ren, "Factors Impacting the Adoption of Artificial Intelligence Powered Cybersecurity Virtual Assistant: A Quantitative Study," Doctoral dissertation, National University, 2024.

[56] M. Pawlicki, M. Chorás, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," *Future Generation Computer Systems*, vol. 110, pp. 148–154, 2020. https://doi.org/10.1016/j.future.2020.11.009

[57] D. Han, Z. Wang, Y. Zhong, W. Chen, J. Yang, S. Lu, and X. Yin, "Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2632–2647, 2021. https://doi.org/10.1109/JSAC.2021.3087242

[58] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. Cordeiro, M. Debbah, and T. H. Luan, "Revolutionizing cyber threat detection with large language models: A privacy-preserving BERT-based lightweight model for IoT/IIoT networks," *IEEE Access*, vol. 12, pp. 23733–23750, 2024. https://doi.org/10.1109/ACCESS.2024.3363973

[59] Z. Xu, D. Li, T. He, and Q. Huang, "Few-shot learning for network intrusion detection using novel feature extraction with autoencoders," *IEEE Transactions on Network and Service Management*, vol. 20, no. 2, pp. 1552–1564, 2023. https://doi.org/10.1109/TNSM.2023.3246009

[60] J. Zhao, S. Shetty, J. W. Pan, C. Kamhoua, and K. Kwiat, "Transfer learning for detecting unknown network attacks," *EURASIP Journal on Information Security*, vol. 1, 1–13, 2019. https://doi.org/10.1186/s13635-019-0084-4

[61] F. Blefari, C. Cosentino, F. A. Pironti, A. Furfaro, and F. Marozzo, "CyberRAG: An Agentic RAG Cyber Attack Classification and Reporting Tool," arXiv:2507.02424, 2025. https://doi.org/10.48550/arXiv.2507.02424

[62] K. Alkhatib and S. Mohammed, "Transformer-based intrusion detection for IoT networks: A comprehensive survey," *IEEE Internet of Things Journal*, vol. 11, no. 6, pp. 9874–9895, 2024.

[63] E. C. Nwosu, O. B. Longe, and N. S. Musa, "Graph-based network intrusion detection using community detection and flow correlation," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 2987–3001, 2022.

[64] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020. https://doi.org/10.1016/j.aiopen.2021.01.001